Formal analytical modeling of blog content as personal narrative

Michael J. Coombs, Holger M. Jaenisch, and James W. Handley

Diplomacy Media Research, Worcester, NY 12197
Alabama A&M University, Department of Physics, Nanoscience Group, Normal, AL 35762
Licht Strahl Engineering INC, Toney, AL 35773

ABSTRACT

This paper contrasts two techniques for analyzing blog content and making use of this information to model blog content. One method uses classical text content and analysis presented for human interpretation. The second method relies on a data mined list of descriptive words characterizing the blogs. We examine the use of different data mining tools, Kryltech’s “Subject Search Summarizer”, Leximancer, and QUEST, to provide orthogonal and independently generated key word lists. These lists are then converted into Data Models, enabling mathematical modeling of blog content.

Keywords: political intelligence, blogs, grounding interpretations, formal analysis tools, Data Modeling

1. BLOGS AS PERSONAL NARRATIVE

The substantive challenges of automating blog analysis are compounded by the ephemeral nature of political blogging. In contrast to mainstream journalism, (a) blogs usually report events from a personal perspective, rather than taking the third-party (although not necessarily non-partisan) perspective of a journalist, (b) blogs mix themes together within a post, rather than exploring a single topic through a well structured narrative, (c) the longevity of themes within blogs often follow no discernable news cycle, and (d) bloggers freely mix passion with objective reporting. To quote from the influential Iraqi blogger Riverbend in her 2006 interview with Al Jazeera:

… I continue because I feel the media covers the situation in my country in a very general way. … Bloggers are not exactly journalists, which is a mistake many people make. They expect us to be dispassionate and unemotional about topics such as occupation and war, etc. That objective lack of emotion is impossible because a blog in itself stems from passion – the need to sit for hours at one’s computer, slouched over the keyboard, trying to communicate ideas, thoughts, fears and frustrations to the world.

In summary, the idiosyncratic nature of blogs seems to provide little by way of classical narrative structure for automated systems to exploit, making it difficult to do more than process the statistics of the lexicon. This contrasts with the structural clarity of traditional narrative forms, including the journalistic forms, which respect the principles of coherence and cohesion in story telling.

For example, an AI system could use a story grammar to look for the grand theme expressed in the reporting of a speech by a leading politician (the unfolding of a country’s history as the journey of a “chosen people”), for a classical plot behind the reporting of a scandal involving the politician (the tragedy of being caught in a blatant sexual lie), or for cause-effect relations in explaining the politician’s behavior (how childhood mistreatment by a priest led eventually to the sexual indiscretion). However, such structure is often absent from blogs, or at best present in a muted form.

If blog sequences could be converted into a formal mathematical time-event line, it then becomes a time series and could be modeled mathematically using the rules and concepts of control theory into a forecastable mathematical model. Such a Data Model is a formal analytical equation of the blog content at an early time point which can be propagated forward as an expectation or prediction of future blog content. This difficult and tricky transformation from abstract ideas encoded as simple key words into a dynamic time line model that encompasses growth and shrinkage of key word variables is the goal of our investigation. We propose an incremental first step in that direction.

*lsei1@yahoo.com; phone 1 256 527 0939.
2. MANUAL TRACING (HUMAN IN THE LOOP) MICROSTORY COMPONENTS OVER MULTIPLE POSTS

2.1. Corpus selection

Baghdad Burning was produced by Riverbend, believed to be a 24-year-old daughter of a former Iraqi diplomat to Washington. Benefiting from her education as part of the Baghdad secular Sunni elite, her life expectations were similar to a Western girl with a similar background. She worked as a computer programmer, and lived free of the constraints on women typical of strict Islamic practice. However, the resurgence of Islamic sentiment in Iraq following the civil chaos that engulfed Iraq early during the occupation began to threaten these freedoms.

The corpus used in our broader study of blog analysis consists of all Riverbend’s posts for August 2003. However, the analysis reported here in Section 2 focuses on the six posts that concern the United States’ (US) attempt to restructure a new Iraqi system of government along the lines of a democracy (posts that deal with the formation of the IGC).

2.2 Analysis

In our speculative case study, an analyst decides to focus on the first month of Riverbend’s blog because a blogger’s core beliefs are often revealed in the early stages of a developing narrative. Figure 1 outlines the content of the six August 2003 posts that deal with the formation of the IGC.

A typical approach to a manual content analysis is to summarize content manually and then to check one’s intuitions against output from an automated system. The software used in this illustration is the Kryltech’s “Subject Search Summarizer”.

- **Monday, Aug. 18, 2003**
  - Jordan advised for Chalabi to be handed over
  - IGC “power hungry freaks” “hand picked” by Bremer
  - Hakim [MRK], Uloom [White cleric], Pechichi, Ciaab, Jaffer [Daewa]
  - A president of an “alphabetical order” as “rule of the month”
  - Result in “Chosen One” selected by Bremer in 9 months in “fake elections”
  - Repeated for best redistributing American assets

- **Tuesday, Aug. 26, 2003**
  - “Rotating presidency failure given ethnicity and Religion as selection criteria”
  - “It's a way of further dividing the Iraqi population”
  - They are puppets of Bremer
  - Jaffer: Gaafar, Shira, violent “activism,” civil centers targets
  - Chalabi: IGC, Sunni, crook, American profile
  - Aslam: IIA, former Baghdadi, Sunni, documents from London
  - Talabani: PUK, Kurd, “escort service,” Berdawani rival
  - Hakim dead, SChIR, Iran, Shura, Badr Brigade commander
  - Pechichi: old, Sunni, foreign minister in Kuwait, emigre
  - Hamad: secretary/columnist, communist party, fundamentalist Sunni group
  - Iloom: old, obscure Shia仇人, London exile
  - Barzani: Kurdistan Democratic party, blood shed in rivalry with Talabani
  - Indicated that Bremer says that rotating presidency represents people
  - America’s puppets, bad start to democracy

- **Friday, Aug. 29, 2003**
  - “Iraq is not a country in chaos” Bremer – Chaos
  - Bremer living in some “other universe”
  - Shooting, locking, abortions of 400 women, executions
  - Mohammed Al-Hakim assassinated, periods by Shia factions
  - Al-Sadr may be involved
  - Riverbend [slander] the guy for what he represented: a puppet and supporter of a fundamentalist Islamic government

- **Saturday, Aug. 30, 2003**
  - No one bothered over killing of Hakim’s brother as “power hungry cleric”
  - People impacted by the additional casualties
  - Iloom resigned from the IGC
  - Bremer should give position to Bush if not re-elected

- **Thursday, Aug. 21, 2003**
  - Chalabi interview with “up close”
  - “Why did the Panther let him?”
  - “Entertainment midget” with IGC militia
  - Chalabi denied he was a “crook”
  - Claimed “IJS, LUB, UIJS”

Figure 1. Summary of Riverbend Posts, August 2003.

- **<<Chalabi is a ‘crook’ wanted in Jordan and yet has been appointed by Bremer as a rotating president of the IGC>>**
- **<<indeed, Bremer has chosen “power-hungry freaks” (and ‘crooks’ by inclusion of Chalabi) “to represent’ the Iraqi people>>**
- **<<Bremer chose the rotation system because he could not decide who of these ‘ingratiating’, ‘dishonest’, ‘incompetent’ people to make president>>**
- **<<moreover, in time Bremer will probably set up ‘fake elections’ to reward the ‘Chosen One’ – the president who best represents ‘American assets in the region’.>>**
Examining the sequence of summaries presented in Figure 1, we conjecture that the analyst will find initial evidence for the emergence of a negative belief cluster on the competence and integrity of the US initiatives to reconstruct the Iraqi government. To do so, the analysis continues by seeking the integrating concepts of a possible belief cluster. The two tools that we consider for reducing small text corpora to their core terms are Leximancer and QUEST. Leximancer employs a Bayesian log-odds metric to abstract concepts and their relations from word co-occurrences in text, and then to display them by mapping to a two-dimensional space. In contrast, QUEST takes word frequency occurrence in documents and maps document similarity into a two-dimensional space where cluster association is performed. Figure 2 gives the Leximancer map for the six August posts, adjusted to contain the names of the principal actors, with the points representing concepts and the circles thematic groupings of concepts. Tags for individual posts are also included to simplify the visualization of concept clustering over time.

Figure 2. Leximancer extraction of 23 concepts for the six IGC-related posts.

Of the 23 concepts extracted by Leximancer, the concepts connecting the most posts are IGC, Bremer, (Al-)Hakim, (Al-)Pachichi, and Sciri. Interestingly, these are all related to stories that express Riverbend’s fear of Iranian and Islamic fundamentalist influence in the new Iraq, along with her contempt for both the Bremer selection of the rotating presidents of the IGC and for the members themselves (e.g., the advanced age of cleric, Al-Pachichi, and the terrorist background of Al-Hakim). There clearly appears to be evidence in the text that theme dominates Riverbend’s debut as a blogger. However, we assume that the analyst wants to proceed further, seeking (a) quantitative evidence for the centrality of this theme and (b) evidence that a negative belief structure is being formed out of the IGC issue. One candidate approach for both structural and quantitative analysis is Formal Concept Analysis (FCA).

In this application, a ‘ Formal Concept’ expresses a relation between a set of concepts, N, and a set of posts, H, where the concepts in N are common to all of the posts in H, and the posts in H satisfy all the concepts in N. This relation can be visualized as a maximal rectangle in an incident matrix of posts and concepts; this is the black rectangle in the concept/post matrix for the Riverbend August posts given in Table 1. Concepts in this table are those defined by the previous Leximancer analysis. Formal Concepts can also be partially ordered by inclusion to form a lattice.
Table 1. Post/concept incidence matrix.

|   | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
| 18|   |   |   |   |   |   |   |   |   |   |   |   | X |   |   |   |   |   |   |   |   |   |   |   |   |
| 21|   |   |   |   |   |   |   |   |   |   |   |   | X |   |   |   |   |   |   |   |   |   |   |   |   |
| 22|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 23|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 26|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 30|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

From these lattices, our analyst not only can identify implicative relations between Leximancer concepts by reading up links but also concepts shared by posts. Common concepts across the corpus include, for example, concerns about the rotating membership of the IGC and the welfare of the Iraqi people. From inspection of these common concepts, our analyst might well conclude that Riverbend fears Shia and Iranian dominance of the IGC under there leadership of the Al-Hakim, along with the fear associated that Iraq could eventually becoming a repressive Islamic state.

Satisfied that the intersection of posts through common Leximancer concepts supports this conclusion, our analyst seeks quantitative conformation. Atkin’s Q-analysis offers one source of measures by computing a graded vector of shared concepts across the posts. Following the Atkin’s convention, vectors are ordered by the number of shared concepts minus one. For the example, the Q-vector is: \( Q = (1^8111221133211111111) \). This indicates that at the higher levels of shared concepts (9 – 12) the lattice divides into three clusters. At the level of ten shared concepts, the three clusters are posts for \{21\}, \{18, 23, 26, 30\}, \{29\}. This can easily be explained by post topic, since the post of August 21 and 29 were stimulated by specific events, namely the Chalabi interview on Jordan’s request that he be repatriated and the assassination of Al-Hakim’s brother respectively. The other four posts are more concerned with Riverbend’s misgivings about the rotating presidents and her fears for the facilitation of fundamentalism and sectarianism in Iraq by members of the IGC. A measure of connective complexity (i.e. fragmentation) can also be computed for this set of posts. This turns out to be a value of 1.55, where total connectivity of all posts would have a value of 1.

Given the small corpus, this result suggests a level of connectivity consistent with a strong negative belief cluster that spans individual posts differentiated by theme.

2.3 Synopsis of manual tracing

Blogs are conceptualized as networks of microstories that are “folksy, without emplotment, a simple telling of chronology;” they also contrasted with mainstream media in mixing passion with fact in reporting personal perspectives on events. Blog narratives tend to be fragmented, blogger developing a story over multiple posts as they are stimulated by ongoing events. We propose that blog analysis should proceed by tracing themes over posts, with actors or core incidents as the carriers. If we than employ formal tools to express the products of analysis, the resulting interpretations are made explicit as they are grounded in the text.

One advantage of such formal grounding is that the basis for an interpretation is laid out for critical comment. Following scientific practice, other analysts may seek to replicate the formal interpretation. Since the formal tools will invariably be replete with parameters, formal modeling will invite exploration of alternative viewpoints. Additionally, formal products may be used to initialize or condition automated analysis systems. In this paper, for example, we have illustrated several ways of converting higher-level story components into simple combinatorial structures that might be used as search keys for one of the standard blog search engines.

A second advantage of formal grounding is that incidence matrix-based representations are common to quantitative approaches to concept analysis. The negative belief structures discussed, for instance, are related to the notion of a constellationary construct as defined by George Kelly’s Personal Construct Theory (PCT). PCT is constructivist psychological theory build around the fundamental postulate that “A person’s processes are psychologically canalized by the ways in which he anticipates events.” Developed by George Kelly in the 1950s for applications to therapeutic discourse, human cognition is viewed as a process of mentally replicating regularities in events through their representation within a system of dichotomous constructs. The theory distinguishes between two classes of entity: elements and constructs. Elements are linguistic entities (nouns, verbs and related grammatical structures) employed to delimit a topic of interest relative to story components such as agents, agency, action, scenes, and purpose. Constructs, on the other hand, define the axes of the coordinate system used to evaluate elements.
A *constellatory construct* in PCT is one that very strongly implies the position of an element on other constructs. This, along with other types of construct employed in PCT analysis, is well defined mathematically. However, in order to benefit from such definition, it would be necessary first to differentiate elements from constructs in our analysis. Our conjecture is that *elements* will be much easier to ground in text than *constructs*. Kelly’s key insight is that constructs require pairs of asymmetric, dichotomous terms for their definition. The asymmetry refers to the observation that, in the naming of an evaluative distinction, one pole is usually made explicit in the process of naming while the other pole is left implicit. In the original setting of clinical psychology, the implicit pole was revealed by various face-to-face questions in which the clinician asks the subject to articulate similarities and differences between elements. This, of course, presents a challenge for intelligence applications to blogs since it is unlikely that an analyst will want to engage directly with the writer.

Our experience with the Riverbend study was that, indeed, elements that were most clearly marked by the formal tools. However, the development of an interpretation certainly benefited from hypothesizing Riverbend’s interpretative dichotomies; in particular, the possible implicit poles. This meant thinking about how Riverbend would have wanted post-invasion Iraq to unfold, where a conjectured “alternative world” for Riverbend’s construal of events becomes another important context to use in evaluating and extending an interpretation.

### 3. MODELING MICROSTORY DEVELOPMENT OVER TIME

If a negative belief cluster is developing, one would expect certain core words to increasingly constrain the content of posts over time. The underlying psychology has to do with what, in its extreme form, is termed ‘obsessive ideation’. The related neural processes are those concerned with the establishment of any social position, where beliefs about the behavior of others are stabilized and acquire emotional color. Indeed, if the Leximancer-extracted concepts are part of a growing belief cluster, one might expect that there will be increasing constraints on the microstories, and, thus, the viewpoints taken in the posts will become more predictable. A method is, therefore, required for measuring the predictability of successive posts in a time series of the full August corpus. This would include, not only the IGC-related posts, but other topics that may include threads of content also found in the IGC microstories.

In this study, we are dealing with a small corpus. It is, therefore, necessary to find a time series analysis technique that is not statistical because of (a) the small number of occurrences of key terms given the small number of posts and (b) the likelihood that salient new term will have a low frequency. QUEST, which is described in detail within the papers given in References 6-8 was originally developed for the real-time analysis of non-indexed Internet sources and works independent of corpus size.

#### 3.1 QUEST

QUEST mines text documents and extracts salient unique words as representation of document concepts. Concept abstraction is achieved by extracting the first occurring sentence containing an identified descriptive word and keeping the sentences in order of occurrence. Separate documents are associated by first mining out a universe of descriptive words across the documents, and then populating a vector equal in length to the descriptive word universe for each document with frequency of occurrence. Descriptive words that occur in more than one blog become concepts, while those that are descriptive but occur in only one blog are called unique words.

Even numbered concepts in the vector are assigned to the y-axis and the odd numbered words in the vector to the x-axis, thereby forming tuples. These tuples form a cloud of points (concepts) from which a centroid for each blog entry is calculated. K-Means clustering is then used on the centroids to group documents that are nearby spatially as documents with similar concepts.

The August blogs entries processed by QUEST were not limited only to the six posts chosen by the analyst that concern the United States’ (US) attempt to restructure a new Iraqi system of government along the lines of a democracy. All of Riverbend’s blogs for August (21 total) were provided as input and QUEST autonomously solved for the concepts and the relationships between blogs. A summary of these blogs entries is provided in Figure 3. Note that the 1st two blog posts on August 17 (17a and 17b) are introductory and general personal narrative about surroundings, therefore they are pulled out separately at the top of Figure 3.
<table>
<thead>
<tr>
<th>17a</th>
<th>Riverbend’s Introduction</th>
<th>17b</th>
<th>Personal narrative of immediate surroundings and 2 different realities</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Jordan’s request that Chalabi be handed over, resolving Ba’athist announcement</td>
<td>24b</td>
<td>65% of Iraqis unemployed, poverty, perineal anger, poverty, and lack of work, fundamentalists killed woman engineer</td>
</tr>
<tr>
<td>19a</td>
<td>Waking up tired because of war stress, dead child killed during US raid, fear of being shot, reminder that under occupation, cannot see future</td>
<td>26a</td>
<td>IEC dividing population into religious and ethnic groups; contempt for the rotating presidents, CPA does not represent Iraqis</td>
</tr>
<tr>
<td>19b</td>
<td>UN building explosion, al-Qaeda claim, occupation renewed fundamentalism, why not guard like Oil Ministry</td>
<td>26b</td>
<td>April 9 a blur of horror, fear of invasion; the national Day has nothing to celebrate</td>
</tr>
<tr>
<td>20</td>
<td>Shock at de Mello’s death, gave hope; terrorism followed occupation, Bremer excuse; US not protecting people; fear if UN not protected</td>
<td>28a</td>
<td>Red Cross withdrawing - expect attacked, who represent detainees? Al-Jazeera, Qamar vs Al-Quds editor, Qamar raids: professionals banned, loyalty to US</td>
</tr>
<tr>
<td>21a</td>
<td>Thanks for emails, variable context, can be creative, can always delete</td>
<td>28b</td>
<td>Iraq sophisticated, educated country before invasion, bridge reconstruction not being given to Iraqi companies, economic rape of Iraq</td>
</tr>
<tr>
<td>21b</td>
<td>New talent at diagnosing gunfire and war machines, introduces brother E</td>
<td>29</td>
<td>Country in chaos since occupation despite Bremer’s beliefs; Hakim’s brother killed, tension among Shia after Iraq–Iran war, backing from Iran/Fundamentalists</td>
</tr>
<tr>
<td>21c</td>
<td>Chalabi interview, “comically frustrating” like Bush, IEC took over country club; car hijacking, said his about Jordan, should ask how spent IEC funds</td>
<td>30a</td>
<td>Talk of assassination, victim not liked; Al-Qaeda responsible caught, prominent SCIRI against Hakim, elderly Doon shaken, resigned, replace with Bush</td>
</tr>
<tr>
<td>22</td>
<td>Do not hate Americans but American troops, personal trauma events, feel terrible for troops in role, WMD to GWOT; bring UN peacekeepers, they are just as young</td>
<td>30b</td>
<td>Gas shortage; family road trip, checkpoints, frustration, bomb damage, looting and killing converted to militia warfare, gangs in Sadr City</td>
</tr>
<tr>
<td>23</td>
<td>Females no longer safe, insults/abduction, education; choice over hijab, violent fundamentalism; SCIRI gave impression that all Shia support; do not, peaceful Iraqs</td>
<td>31</td>
<td>WMD joke</td>
</tr>
<tr>
<td>24a</td>
<td>About Riverbend, raised abroad, bilingual, many Iraqis know all about Western culture; remain anonymous to be safe</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Summary of the August posts. Blog entries (posts) on August 17 (17a and 17b) are introductory and general narrative about surroundings, therefore they are pulled out separately at the top.

In mining the twenty one August blog entries QUEST found 1576 candidate concepts across the 21 posts (full description of how QUEST identifies candidate concepts is found in Reference 13). Once the candidate concepts are identified, QUEST identified the number of posts that contained each unique word by converting the frequency of occurrence vector into a 0/1 mask and summing across the number of entries. Using this summed list, the concepts that exist in the posts were found by keeping all of the unique words that occurred in 2 or more posts. A total of 775 concepts were found across the 21 entries. The remainder is labeled as unique words.

The QUEST clustering results for the 21 blog entries after selecting all of the descriptive words as concepts are shown in Figure 4, with three groupings labeled as A, B, and C. In Figure 4, the further up the diagonal towards the right a post falls, the more of the total concepts it embodies. The polar opposite at the bottom left corner, where the furthest entry represents virtually no concept content in the blog entry. Specifically, in Group A blog 24b is a single post that best encompasses all of the concepts and topics embodied by the entire corpus and may be thought of as a corpus of the abstract. In contrast, its polar opposite, post 31 contains virtually none of the topics or concepts contained in the corpus and would be most distinctly and markedly different from other posts in the corpus. The first post, entry 24b in summary discusses unemployment, poverty, anger, and frustration, in general all of the problems and environmental problems that directly impact the life of Riverbend. The final post, entry 31 merely suggests a link to a joke about weapons of mass destruction (WMD).

Figure 5 shows the results for choosing the same concepts as the ones previously identified by Leximancer (IGC was not identified by QUEST as a concept; therefore, it could not be selected). The grouping of points near the lower left (Others) are the documents in the blog corpus that do not contain any of the selected concepts. The remaining two groups are posts that have selected concepts in common, but are grouped into two groups which indicate that the topics of discussion or philosophy would be distinctly between the two groups, the exception being post 29 which overlaps both groups and is probably a good summary or abstract post for both groups.
Figure 4. QUEST clustering and grouping using all of the concepts identified in the blog entries. A total of 3 groups were identified by QUEST, which are labeled as A, B, and C.

Figure 5. QUEST clustering and grouping using the Leximancer identified concepts identified as connecting the most posts (Note that IGC was not identified by QUEST as a concept due to its length).

QUEST identified a total of eight blogs entries (18, 19b, 20, 21c, 23, 26a, 29, 30b) as being concept related. It was shown in Section 2.2 that Leximancer identified six posts as concept related, all of which were also identified by QUEST. QUEST also identified two additional blogs (19b and 20). Upon inspection by the analyst, it was found that these additional posts also contained the concepts identified by Leximancer and likely should have been included as part of the original Leximancer grouping. Also in Section 2.2, Atkins Q-analysis divided the lattice of shared components (concepts) into three blog clusters. The grouping of the eight posts by QUEST separated the posts into two clusters and
grouped them differently than the Q-analysis as seen in Figure 5. We now turn out attention to the idea of predicting changes in post topic.

3.2 Formal Modeling, Analysis, and Prediction
To start the process of modeling blog content, we begin with the 775 concepts that were identified in the previous section for the blog corpus. These concepts occurred in at least two of the blog entries, which leads to the question “How many posts have a fixed number of concepts in common?” In order to answer this question, we used the QUEST frequency of occurrence file that is generated with the run. This file contains a matrix, where the rows are individual posts and the columns are different concepts. By summing down the matrix columns (across different posts), we determined the number of entries that contained each of the concepts and we calculated standard statistics on the number of blog entries such as the mean, median, and mode. The number of posts was then sorted into descending order, and each of the concepts was also sorted along with the number of posts simultaneously. This produced the data shown on Figure 6 as the dashed curve.

![Figure 6. QUEST derived number of blogs versus concepts in common. Data derived from the blog entries is shown as the dashed line, while the Data Model given in Equation (1) is shown as the solid line.](image)

Next, a polynomial Data Model was derived that models the number of blog entries versus the fixed number of concepts in common, which is denoted by the variable $C$ and given by the mathematical expression

$$\text{Blogs}(C) = 10.541 - 0.056C + 1.516 \cdot 10^{-4} C^2 - 2.007 \cdot 10^{-7} C^3 + 1.235 \cdot 10^{-10} C^4 - 2.829 \cdot 10^{-14} C^5$$

and shown in Figure 6 as a solid line. To eliminate the need to solve Equation (1) for $C$, we derived a second Data Model for the number of fixed concepts in common versus the number of blog entries (here given by $B$) which is given by the mathematical expression

$$\text{Concepts}(B) = 2.133 \cdot 10^3 - 1.225 \cdot 10^3 B + 299.179 B^2 - 35.788 B^3 + 2.029 B^4 - 0.043 B^5$$

Using the median number (8) of posts found previously when calculating the standard statistics as input into Equation (2), we determine that the number of dominant concepts is 46. Using the concepts from the QUEST output, we keep the 46 concepts that correspond to the 46 most frequently occurring concepts in the blog corpus, calling them the dominant concepts. We then list below for each of the posts the dominant concepts in each.
August 17a – about, think, Iraqi, could, matters, start, that's, these
August 17b – about, Americans, trying, reality, something, around, maybe, these
August 18 – people, about, think, Americans, Iraqi, would, could, occupation, council, trying, don't, before, Iraqis, decide, reality, Bremer, around, house, whole, someone, found, another, anyone, which, during, that's, these
August 19a – people, about, think, Americans, Iraqi, occupation, Baghdad, don't, tanks, Iraqis, reality, can't, matters, maybe, house, where, someone, found, anyone, during, that's, family
August 19b – people, think, Iraqi, would, could, occupation, Baghdad, council, tanks, before, being, Bremer, something, believe, supposed, couldn't, these
August 20 – people, think, Americans, Iraqi, would, could, occupation, Baghdad, trying, Iraqis, Bremer, matters, someone, couldn't, going, during, that's, wasn't, inside, these
August 21a – people, about, would, trying, don't, tanks, being, reality
August 21b – people, about, think, Americans, would, could, council, trying, don't, decide, reality, can't, something, maybe, isn't, house, where, whole, supposed, found, anyone, which, that's, wasn't, front, inside
August 21c – people, about, think, Iraqi, would, could, Baghdad, Iraqis, decide, reality, Bremer, can't, again, start, whole, someone, another, during, wasn't,
August 22 – people, about, think, Americans, Iraqi, could, occupation, Baghdad, council, trying, don't, tanks, before, being, Iraqis, decide, can't, again, something, believe, around, maybe, isn't, house, start, whole, someone, couldn't, found, anyone, going, front, inside, family
August 23 – people, about, think, Americans, Iraqi, would, could, occupation, Baghdad, council, don't, before, being, Iraqis, decide, reality, can't, again, something, believe, matters, around, isn't, house, whole, someone, supposed, another, going, which, during, that's, front, inside, family
August 24a – people, about, think, Americans, Iraqi, could, Baghdad, reality, believe, matters, maybe, which,
August 24b – people, about, think, Americans, Iraqi, would, could, occupation, Baghdad, trying, don't, tanks, before, being, reality, Bremer, again, something, believe, matters, around, maybe, isn't, house, where, start, whole, someone, supposed, couldn't, found, anyone, going, which, during, wasn't, front, family, these
August 26a – people, about, think, Americans, Iraqi, would, occupation, Baghdad, council, tanks, before, being, Iraqis, decide, reality, Bremer, isn't, house, where, start, supposed, couldn't, found, anyone, going, inside, these,
August 26b – Americans, would, could, occupation, Baghdad, council, trying, tanks, before, Bremer, can't, something, matters, around, maybe, house, where, start, anyone, front
August 28a – people, about, think, Americans, Iraqi, would, could, occupation, council, trying, don't, tanks, before, being, Iraqis, decide, can't, again, something, believe, matters, around, isn't, where, start, whole, someone, supposed, couldn't, found, another, which, during, that's, wasn't, front, inside, family
August 28b – people, about, think, Americans, Iraqi, would, could, occupation, council, trying, don't, tanks, before, being, Iraqis, decide, can't, again, something, believe, matters, around, isn't, where, start, whole, someone, supposed, couldn't, found, another, which, during, that's, wasn't, front, inside, family
August 29 – people, about, think, Americans, occupation, Baghdad, council, don't, tanks, being, Bremer, again, believe, maybe, where, another, going, wasn't, family
August 30a – people, Americans, Iraqi, would, council, decide, Bremer, again, believe, maybe, isn't, supposed, found, going
August 30b – people, about, think, Americans, Iraqi, would, could, occupation, Baghdad, trying, don't, tanks, before, being, Iraqis, decide, can't, again, around, isn't, house, where, start, whole, another, anyone, going, which, wasn't, front, inside, family
August 31 – search, access

4. SUMMARY

Blogs are conceptualized as networks of microstories that are a simple telling of chronology and contrasted with mainstream media in mixing passion with fact in reporting personal perspectives on events. Blog narratives tend to be fragmented, with a blogger developing a story over multiple posts as they are stimulated by ongoing events. We propose that blog analysis should proceed by tracing themes over posts, with actors or core incidents as the carriers. If we than convert the post sequences into a formal mathematical time-event line. And if post sequences can be converted into a formal mathematical time-event line, it then becomes a time series and could be modeled mathematically using the rules and concepts of control theory into a mathematical time-event line. And if post sequences can be converted into a formal mathematical time-event line, it then becomes a time series and could be modeled mathematically using the rules and concepts of control theory into a

Using the six IGC related blogs, we performed classical text content and analysis using Leximancer and Kryltech’s “Subject Search Summarizer” to extract information for the human analyst to interpret. We also demonstrated in this paper that blog sequence content can be automatically mined and their content analyzed. This was done using QUEST, which mined the descriptive words from the blog entry (post) sequences autonomously and used frequency of occurrence to analyze the descriptive words and converted the list into concepts (common to two or more posts) and unique words (occurring in only one post).

An analytical Data Model was derived that models blog content for the number of posts that have a fixed number of concepts in common. This demonstration serves as a first step to converting the post sequences into a formal mathematical time-event line. And if post sequences can be converted into a formal mathematical time-event line, it then becomes a time series and could be modeled mathematically using the rules and concepts of control theory into a
forecastable mathematical Data Model. Such a Data Model is a formal analytical equation of the blog content at an early time point which can be propagated forward as an expectation or prediction of future blog content.

This difficult and tricky transformation from abstract ideas encoded as simple key words into a dynamic time line model that encompasses growth and shrinkage of key word variables is the goal of our ongoing research, which continues in a companion paper to this one (Reference 13), where we convert the blog content into a time series and derive an analytical Data Model that can be used both to model blog content and predict future blog content.

ACKNOWLEDGEMENTS

The authors would like to thank Jeff Faucheux for his analysis ideas and for providing the Internet links for data used in this work.

REFERENCES

1. http://riverbendblog.blogspot.com